Efficient Universal Noiseless Source Codes

LEE D. DAVISSON, FELLOW, IEEE, ROBERT J. McELIECE, MEMBER, IEEE, MICHAEL B. PURSLEY, SENIOR MEMBER, IEEE, AND MARK S. WALLACE

Abstract—**Although the existence of universal noiseless variable-rate codes for the class of discrete stationary ergodic sources has previously been established, very few practical universal encoding methods are available. Efficient implementable universal source coding techniques are discussed in this paper. Results are presented on source codes for which a small value of the maximum redundancy is achieved with a relatively short block length. A constructive proof of the existence of universal noiseless codes for discrete stationary sources is first presented. The proof is shown to provide a method for obtaining efficient universal noiseless variable-rate codes for various classes of sources. For memoryless sources, upper and lower bounds are obtained for the minimax redundancy as a function of the block length of the code. Several techniques for constructing universal noiseless source codes for memoryless sources are presented and their redundancies are compared with the bounds. Consideration is given to possible applications to data compression for certain nonstationary sources.**

I. Introduction

THE PROBLEM of variable-rate source coding for sources with unknown or incompletely specified probability distributions has received considerable attention in the recent literature (e.g., [1]–[10]). The problem is to find a single variable-rate code which is optimum for each source in a given class. A sequence of variable-rate codes of increasing block lengths which is asymptotically optimum as measured by redundancy is said to be universal for the given class of sources. Several previous papers have dealt with the existence of such sequences of codes (e.g., [1], [3], [6], and [9]). However, these are primarily asymptotic results which do not give an indication of the block length required to achieve a given redundancy or how to actually obtain a code which will achieve a given redundancy. In the present paper, both of these problems are addressed.

In Section II we present an alternative proof of the existence of universal noiseless source codes for an arbitrary class of discrete stationary sources. The interest in this proof is primarily due to the insight that it provides for the construction of universal source codes. The technique employed in the proof actually gives a universal code construction, the performance of which is evaluated in this paper.

Manuscript received December 7, 1979; revised July 2, 1980. This work was supported by the National Science Foundation under Grants ENG75-20864 at the University of Illinois, Urbana, IL 61801, and ENG77-10503 at the University of Maryland, College Park, MD 20742, and by the Joint Services Electronics Program at the University of Illinois under Contract N00014-79-C-0424.

L. D. Davisson is with the Department of Electrical Engineering, University of Maryland, College Park, MD 20742.

R. J. McEliece, M. B. Pursley, and M. S. Wallace are with the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801.

In Section III we derive new upper and lower bounds on the minimax redundancy as a function of the block length of the code. A family of lower bounds is obtained via a rate-distortion-theoretic argument which is quite general. The basic lower bounds are obtained for an arbitrary class of discrete stationary sources, and the results are then specialized to memoryless sources. The upper bound is obtained by the construction of a specific code.

In Section IV the technique presented in Section II is employed to obtain a method for the construction of universal source codes for the class of binary memoryless sources. The redundancy of this scheme is evaluated for short to moderate block lengths.

In Section V several modifications are made to the basic code construction method of Section IV. These modifications give smaller values of maximum redundancy for block lengths of interest. The redundancies of the codes obtained are compared with the bounds of Section III. Two important constraints on the code block lengths are as follows. The constraint which is the most severe in practice is the constraint imposed by the necessity to limit the complexity of the encoding and decoding algorithms. A second constraint arises in many practical applications where the source output probability distribution may be changing slowly. These two constraints impose quite different restrictions on the parameters of the codes introduced in Section V. In Section V we also present results on two alternative coding techniques which have been presented in the literature. Comparisons are made between these codes and the codes that are obtained from our constructions.

II. Existence of Universal Noiseless Source Codes

Let A be a finite set, and for each positive integer n let A^n denote the set of all n-vectors $x=(x_1, x_2, \cdots, x_n)$ with elements x_i from A. By a class or collection of discrete stationary sources we mean a subset of the set of all stationary sources with the given finite alphabet A. Let Λ be an index set corresponding to a class of discrete stationary sources. Each $\theta \in \Lambda$ corresponds to a particular source output distribution which is characterized by the probability $p_n(x|\theta)$ of the source output x from the θth source. Let X be a random n-vector from the source, and for an arbitrary real-valued function f defined on A^n let

$$E_\theta\{f(X)\} = \sum_x f(x) p_n(x|\theta).$$

A binary variable-length source code of order n is a mapping b_n from A^n into the set of all finite-length binary

0018-9448/81/0500-0269$00.75

sequences. For our purposes, a code b_n is completely characterized by its associated length function l_n, which is defined by setting $l_n(x)$ equal to the length of the codeword $b_n(x)$. A necessary and sufficient condition for the function l_n to correspond to at least one uniquely decodable binary variable-length code is the Kraft inequality

$$\sum_x 2^{-l_n(x)} \leq 1.$$

Let $\mathcal{K}_n$ denote the set of all functions l_n which satisfy this inequality.

A brief review of the concepts and terminology of universal source coding [1] may be of help to the reader, and it will also serve to introduce the notation to be used throughout the paper. Given any code with codeword lengths specified by the function l_n, the average codeword length that results when the code is applied to the source θ is given by

$$\bar{l}_n(\theta) = E_\theta\{l_n(X)\} = \sum_x l_n(x) p_n(x|\theta). \tag{1}$$

Let $H_n(\theta)$ denote the nth order per-letter entropy[1] of the source θ; that is,

$$H_n(\theta) = -n^{-1} \sum_x p_n(x|\theta) \log p_n(x|\theta). \tag{2}$$

Since $n^{-1}\bar{l}_n(\theta) \geq H_n(\theta)$, the nth order redundancy

$$r_n(l_n, \theta) = n^{-1}\bar{l}_n(\theta) - H_n(\theta) \tag{3}$$

for the code l_n applied to the source θ is nonnegative. For a given code l_n let

$$\hat{r}_n(l_n) = \sup\{r_n(l_n, \theta): \theta \in \Lambda\}, \tag{4}$$

so that the *nth order minimax redundancy* for the class Λ is given by

$$\mathcal{R}_n = \inf\{\hat{r}_n(l_n): l_n \in \mathcal{K}_n\}. \tag{5a}$$

The quantity $\mathcal{R}_n$ is a per-letter redundancy. For convenience we also define an unnormalized redundancy $\mathcal{R}_n^*$ by

$$\mathcal{R}_n^* = n\mathcal{R}_n. \tag{5b}$$

A sequence $(l_n) = l_1, l_2, l_3, \cdots$ of variable-length source codes is *weakly minimax universal* [1] or *weakly universal* [6] for the class Λ if

$$\lim_{n\to\infty} r_n(l_n, \theta) = 0, \qquad \text{for all } \theta \in \Lambda, \tag{6}$$

and it is *strongly universal* [6] if

$$\lim_{n\to\infty} n^{-1}\bar{l}_n(\theta) = H(\theta) \qquad \text{[uniformly]}, \tag{7}$$

where $H(\theta)$ is the entropy of the discrete stationary source θ. In view of (3) and the fact that

$$\lim_{n\to\infty} H_n(\theta) = H(\theta),$$

it is clear that (7) implies (6), but (7) is a much stronger property since it implies uniform convergence of $r_n(l_n, \theta)$ rather than the weaker pointwise convergence of (6). If the sequence $\{l_n\}$ is strongly universal, then it is also *minimax universal* [1]; that is, (7) implies $\lim_{n\to\infty} \hat{r}_n(l_n) = 0$.

The following result is useful in establishing the existence of universal source codes and in the construction of such codes.

Lemma 1: Let Λ represent a class of discrete stationary sources with common finite alphabet A. For each postive integer n there exists a collection $\{b_n^{(k)}: 1 \leq k \leq K_n\}$ of binary variable-length codes of order n and a collection $\{\Lambda_n^{(k)}: 1 \leq k \leq K_n\}$ of subsets of Λ such that

$$\Lambda = \bigcup_{k=1}^{K_n} \Lambda_n^{(k)}, \tag{8}$$

and

$$n^{-1}E_\theta\{l_n^{(k)}(X)\} < H_n(\theta) + 2n^{-1}, \tag{9}$$

for each $\theta \in \Lambda_n^{(k)}$ where $l_n^{(k)} \in \mathcal{K}_n$ is the length function for the code $b_n^{(k)}$.

Proof: Let $\mathbb{P}(A^n)$ be the set of all discrete density functions (probability mass functions) on A^n. Clearly, $\mathbb{P}(A^n)$ is a compact convex subset of the normed linear space consisting of all real-valued functions on A^n with the norm

$$\|f\| = \max\{|f(x)|: x \in A^n\}. \tag{10}$$

Let m be an arbitrary positive integer and let $\delta = (m+1)^{-1}J^{-2n}$ where $J = |A|$ is the cardinality of A. Let Q_δ be the set of all $q \in \mathbb{P}(A^n)$ such that for each $x \in A^n$, $q(x) = i\delta$ for some integer $i > 0$. Notice that Q_δ is finite (in fact $\log|Q_\delta| < -J^n \log \delta$) and that for each $p \in \mathbb{P}(A^n)$ there is at least one $q \in Q_\delta$ such that $\|p - q\| \leq (J^n - 1)\delta$. Notice that if p is such that $p(x_0) = 1$, then there is exactly one such q (namely, $q(x) = \delta$ for $x \neq x_0$ and $q(x_0) = 1 - (J^n - 1)\delta$). Now consider the per-letter relative entropy (or divergence) which is defined by

$$H_n(p, q) = n^{-1} \sum_x p(x) \log[p(x)/q(x)], \tag{11}$$

where, as usual, $0\log 0 = 0$ in such expressions. We are interested in $H_n(p, q)$ only for $q \in Q_\delta$ so that $q(x) \geq \delta$ for all $x \in A^n$. Define

$$f_n(p, \delta) = \min\{H_n(p, q): q \in Q_\delta\}.$$

Given $p \in \mathbb{P}(A^n)$, define $\hat{q}$ as follows. First order A^n such that $p(x_1) \geq p(x_2) \geq \cdots \geq p(x_M)$, where $M = J^n$. Notice that $p(x_1) \geq J^{-n}$. For each j in the range $2 \leq j \leq M$, define

$$\hat{q}(x_j) = \lfloor 1 + \delta^{-1}p(x_j) \rfloor \delta,$$

where $\lfloor u \rfloor$ denotes the integer part of the real number u. It follows that

$$\sum_{j=2}^{M} p(x_j) < \sum_{j=2}^{M} \hat{q}(x_j) \leq (M-1)\delta + \sum_{j=2}^{M} p(x_j),$$

and so if we let

$$\hat{q}(x_1) = 1 - \sum_{j=2}^{M} \hat{q}(x_j),$$

[1] Base two logarithms are used throughout the paper.

then

$$0<p(x_1)-\hat{q}(x_1)\leq(M-1)\delta. \tag{12}$$

It follows that $\hat{q}\in Q_\delta$ and thus

$$f_n(p,\delta)\leq H_n(p,\hat{q})<n^{-1}p(x_1)\log[p(x_1)/\hat{q}(x_1)], \tag{13}$$

since $\log[p(x_j)/\hat{q}(x_j)]<0$ for $2\leq j\leq M$. But from (12) we see that

$$\frac{\hat{q}(x_1)}{p(x_1)}\geq 1-\frac{(M-1)\delta}{p(x_1)}\geq 1-M(M-1)\delta$$
$$\geq\frac{m}{m+1}\geq\frac{1}{2}.$$

Therefore

$$f_n(p,\delta)<n^{-1}p(x_1)\log m^{-1}(m+1)$$
$$\leq n^{-1}. \tag{14}$$

Let $\{q^{(k)}\colon 1\leq k\leq K_n\}$ be an ordering of the set Q_δ (e.g., a lexicographic ordering will suffice) and define

$$\Lambda_n^{(k)}=\{\theta\in\Lambda\colon H_n(p_\theta,q^{(k)})<n^{-1}\}, \tag{15}$$

where p_θ denotes the function defined by $p_\theta(x)=p_n(x|\theta)$. In view of (12) and (14) we see that for each $\theta\in\Lambda$ there is a k such that $H_n(p_\theta,q^{(k)})<n^{-1}$. Hence the sets $\Lambda_n^{(k)}$ of (15) satisfy (8).

Next let $b_n^{(k)}$ be the Shannon code for $q^{(k)}$ so that

$$l_n^{(k)}(x)=\lceil-\log q^{(k)}(x)\rceil \tag{16}$$

where $\lceil u\rceil$ denotes the smallest integer greater than or equal to the real number u. If this code is applied to the source θ then the average codeword length is

$$n^{-1}E_\theta\{l_n^{(k)}(X)\}\leq n^{-1}\sum_x\left[1-\log q^{(k)}(x)\right]p_n(x|\theta)$$
$$=n^{-1}\left[1-\sum_x p_n(x|\theta)\log q^{(k)}(x)\right]$$
$$=n^{-1}+H_n(p_\theta,q^{(k)})+H_n(\theta). \tag{17}$$

Because of (15), (17) implies that (9) holds for all $\theta\in\Lambda_n^{(k)}$, thus completing the proof of Lemma 1.

The following result is a slight improvement of [1, thm. 7] in that it applies to an arbitrary class of discrete stationary sources (the sources need not be ergodic). However the most significant contribution of this section is not the theorem itself. The proof of the theorem is new, and it is based on Lemma 1 rather than on histogram encoding as in [1]. Not only does this provide an alternative proof of the existence of universal noiseless source codes, but more importantly it gives a universal code construction that is developed further in a later section where it is shown to be quite efficient.

Theorem 1: Weakly universal variable-rate codes exist for an arbitrary collection of discrete stationary sources. Strongly universal variable-rate codes exist for a given class Λ of discrete stationary sources if and only if the nth order entropy $H_n(\theta)$ converges to the entropy $H(\theta)$ *uniformly* on Λ.

Remark: For example, the uniform convergence condition is satisfied if for some k the class of sources is contained in the set of all kth order Markov sources. It is not satisfied for the class of all discrete stationary sources or even the class of all stationary Markov sources.

Proof: For each n apply Lemma 1 to obtain a collection $\{b_n^{(k)}\colon 1\leq k\leq K_n\}$ of codes and a collection $\{\Lambda_n^{(k)}\colon 1\leq k\leq K_n\}$ of subsets of Λ for which (8) holds and

$$E_\theta\{l_n^{(k)}(X)\}\leq nH_n(\theta)+2, \tag{18}$$

for each $\theta\in\Lambda_n^{(k)}$. For each integer n define

$$L(n)=\max\{n,\lceil\log K_n\rceil\}, \tag{19}$$

and let $M(n)=L(n+1)-L(n)$. For a given n apply the code $\{b_n^{(k)}\colon 1\leq k\leq K_n\}$ to source words of length $n[L(n)+m]$ where $0\leq m<M(n)$. For such an integer m let $L=L(n)+m$ and $N=nL$, and consider the encoding of a source word y of length N. Segment y into $L(n)+m$ blocks of length n, and write $y=(x_1,x_2,\cdots,x_L)$ where x_j is the segment of n consecutive source elements beginning with x_{jn-n+1} and ending with x_{jn}. Encode y by applying the code $b_n^{(k')}$ to each of the segments x_j where k' is such that

$$\sum_{j=1}^{L}l_n^{(k')}(x_j)=\min\left\{\sum_{j=1}^{L}l_n^{(k)}(x_j)\colon 1\leq k\leq K_n\right\}. \tag{20}$$

In order to make this encoding method uniquely decodable, add the $\lceil\log K_n\rceil$-bit binary representation of the integer k' to the beginning of the codeword for y. Applying this procedure for each N-tuple y, we obtain a code b_N with codeword lengths

$$l_N(y)=\lceil\log K_n\rceil+\min\left\{\sum_{j=1}^{L}l_n^{(k)}(x_j)\colon 1\leq k\leq K_n\right\}. \tag{21}$$

From (21) and (18) it follows that for each $\theta\in\Lambda_n^{(k)}$

$$E_\theta\{l_N(Y)\}\leq\lceil\log K_n\rceil+\sum_{j=1}^{L}E_\theta\{l_n^{(k)}(X_j)\}$$
$$\leq L(n)+L\{nH_n(\theta)+2\}$$
$$\leq N\{H_n(\theta)+3n^{-1}\}. \tag{22}$$

Thus, for each $\theta\in\Lambda$

$$H(\theta)\leq H_N(\theta)\leq N^{-1}E_\theta\{l_N(Y)\}\leq H_n(\theta)+3n^{-1}, \tag{23}$$

which holds for $N\in\{n[L(n)+m]\colon 0\leq m<M(n)\}$. This procedure gives a sequence of codes of orders $\{n[L(n)+m]\colon n=1,2,3,\cdots;0\leq m<M(n)\}$ where $M(n)=L(n+1)-L(n)$ and $L(n)$ is given by (19). In order to get a full

sequence of codes (i.e., of orders $1,2,3,\cdots$) it is necessary to fill in the "gaps" between $n[L(n)+m]$ and $n[L(n)+m+1]$. Suppose that N' is such that

$$N = n[L(n)+m] < N' < n[L(n)+m+1]. \quad (24)$$

A code $b_{N'}$ for source words of length N' can be obtained by applying the code b_N to the first N digits of the source words. This leaves at most $n-1$ source digits to be encoded. But these can be encoded using $(n-1)\log J$ bits, where J is the size of the source alphabet A. Hence the new code has average codeword length

$$E_\theta\{l_{N'}(Y)\} \leq N\{H_n(\theta)+3n^{-1}\} + (n-1)\log J, \quad (25)$$

and so (23) becomes

$$\begin{aligned} H(\theta) &\leq (N')^{-1} E_\theta\{l_{N'}(Y)\} \\ &\leq H_n(\theta) + 3n^{-1} + (N')^{-1}(n-1)\log J \\ &\leq H_n(\theta) + 3n^{-1} + [L(n)]^{-1}\log J \\ &\leq H_n(\theta) + (3+\log J)n^{-1}. \end{aligned} \quad (26)$$

The fact that the resulting sequence is weakly universal follows from (26) and the observation that (24) implies $N' \to \infty$ if and only if $n \to \infty$. Thus we have

$$H(\theta) \leq \lim_{N'\to\infty} (N')^{-1} E_\theta\{l_{N'}(Y)\} \leq \lim_{n\to\infty} H_n(\theta) = H(\theta). \quad (27)$$

Notice from (27) that the sequence is strongly universal if

$$\lim_{n\to\infty} H_n(\theta) = H(\theta) \qquad \text{[uniformly]}. \quad (28)$$

Therefore all that remains to be proved is the necessity of (28) for strongly universal codes, but this follows immediately from the fact that $n^{-1}\bar{l}_n(\theta) \geq H_n(\theta) \geq H(\theta)$. If $H_n(\theta)$ does not converge uniformly to $H(\theta)$, then it is impossible for $n^{-1}\bar{l}_n(\theta)$ to converge uniformly to $H(\theta)$.

III. Bounds on the Redundancy of Universal Noiseless Source Codes

A. Lower Bounds

Suppose the unknown parameter θ is modeled as a random variable Θ which takes the values in the set Λ. Then the nth order redundancy of (3) becomes the random variable $r_n(l_n,\Theta)$ for a given code l_n. From (1)–(3) we see that this random variable has the expected value

$$\begin{aligned} &E\{r_n(l_n,\Theta)\} \\ &= n^{-1}\left[\sum_x p_n(x) l_n(x) - E\left\{-\sum_x p_n(x|\Theta)\log p_n(x|\Theta)\right\}\right] \\ &= n^{-1}\left[\sum_x p_n(x) l_n(x) - H(X|\Theta)\right]. \end{aligned} \quad (29)$$

In (29) X is the random n-vector representing the source output, p_n is the discrete density function defined by

$$p_n(x) = E\{p_n(x|\Theta)\} \quad (30)$$

for each $x \in A^n$, and $H(X|\Theta)$ is the nth order conditional entropy of the source output given Θ. The quantity $E\{r_n(l_n,\Theta)\}$ is known as the *average redundancy* [1] for l_n. Since

$$\sum_x p_n(x) l_n(x) \geq H(X) \triangleq -\sum_x p_n(x)\log p_n(x), \quad (31)$$

then

$$\begin{aligned} nE\{r_n(l_n,\Theta)\} &\geq H(X) - H(X|\Theta) \\ &= I(X;\Theta) \end{aligned} \quad (32)$$

where $I(X;\Theta)$ is the mutual information between the source output and the parameter. Let $\hat{\Theta}$ be any random variable for which $\Theta \to X \to \hat{\Theta}$ forms a Markov chain (i.e., Θ and $\hat{\Theta}$ are conditionally independent given X). Let $\hat{\Lambda}$ denote the set of values taken on by $\hat{\Theta}$ and consider an arbitrary function d: $\Lambda \times \hat{\Lambda} \to [0,\infty)$. Consider the rate-distortion function for the estimation of Θ by $\hat{\Theta}$ using fidelity criterion d. If

$$D \triangleq E\{d(\Theta,\hat{\Theta})\}, \quad (33)$$

then the definition of the rate-distortion function and the fact that $\Theta \to X \to \hat{\Theta}$ is a Markov chain imply

$$R(D) \leq I(\Theta;\hat{\Theta}) \leq I(\Theta;X). \quad (34)$$

From (32) and (34) we see that

$$nE\{r_n(l_n,\Theta)\} \geq R(D). \quad (35)$$

The inequality in (35) gives a family of lower bounds on the average redundancy. A bound is obtained for each choice of the conditional distribution of $\hat{\Theta}$ given X and each choice of the distortion measure d. Of greater interest for our purposes is the fact that (35) also gives a lower bound on the nth order minimax redundancy $\mathcal{R}_n$ defined in (5). This follows from the fact that for any $l_n \in \mathcal{K}_n$

$$\begin{aligned} \hat{r}_n(l_n) &\triangleq \sup\{r_n(l_n,\theta): \theta \in \Lambda\} \geq E\{r_n(l_n,\Theta)\} \\ &\geq n^{-1}R(D), \end{aligned} \quad (36)$$

and therefore

$$n^{-1}\mathcal{R}_n^* = \mathcal{R}_n \triangleq \inf\{\hat{r}_n(l_n): l_n \in \mathcal{K}_n\} \geq n^{-1}R(D), \quad (37)$$

It is important to notice that (37) holds for all probability distributions for the random variable Θ. Consequently (37) and (35) provide a family of lower bounds on $\mathcal{R}_n$. A lower bound is obtained for each choice of the distortion measure d, the distribution of Θ, and the conditional distribution of $\hat{\Theta}$ given X. This result is summarized in the following theorem.

Theorem 2: The nth order minimax redundancy $\mathcal{R}_n$ for a class Λ of discrete stationary sources satisfies $\mathcal{R}_n \geq n^{-1}\sup R(D)$ where $R(D)$ is the rate-distortion function for reproduction of Θ by $\hat{\Theta}$ under distortion measure d and where the supremum is over all distortion measures and all

probability distributions for $(\Theta, X, \hat{\Theta})$ for which $\Theta \to X \to \hat{\Theta}$ is a Markov chain and $P(X = x|\Theta = \theta) = p_n(x|\theta)$.

The condition that $\Theta \to X \to \hat{\Theta}$ is a Markov chain is automatically satisfied if $\hat{\Theta} = g(X)$ for a function g: $A^n \to \hat{\Lambda}$, which is the situation for the application of Theorem 2 in this present paper. In this case $\hat{\Theta}$ represents an estimate of the parameter Θ which is a (deterministic) function of the source output X. This is an intuitively satisfying notion in the context of the universal source coding where the encoder sees X but does not know the value of Θ.

It would appear that for many applications of Theorem 2—and this is certainly the case for the application in this paper—it is desirable to have a lower bound which is simpler and easier to compute than the rate-distortion lower bound of Theorem 2. This leads us to the consideration of the Shannon lower bound $R_L(D)$ for the rate-distortion function $R(D)$. We consider now the case $\Lambda = \hat{\Lambda}$ where Λ is the real line $\mathbb{R}$, or more generally a subset of m-dimensional Euclidean space $\mathbb{R}^m$. We restrict attention to difference distortion measures $d(\theta, \hat{\theta}) = \rho(\theta - \hat{\theta})$ and continuous distributions for Θ. The Shannon lower bound applies in this situation so that we may write $\mathcal{R}_n \geq n^{-1} R_L(D)$ where $R_L(D)$ is the Shannon lower bound for density w, difference distortion measure ρ, and any conditional distribution of $\hat{\Theta}$ given X. In particular, if

$$d(\theta, \hat{\theta}) = \rho(\theta - \hat{\theta}) = \|\theta - \hat{\theta}\|^2, \tag{38}$$

where $\|u\| = [u_1^2 + u_2^2 + \cdots + u_m^2]^{1/2}$ is the usual norm on $\mathbb{R}^m$, then if Λ has dimension m the Shannon lower bound is

$$R_L(D) = h(w) - \tfrac{1}{2} m \log 2\pi e D, \tag{39}$$

where $h(w)$ is the differential entropy

$$h(w) = -\int_\Lambda w(\theta) \log w(\theta)\, d\theta. \tag{40}$$

Suppose that Λ represents the class of all memoryless sources with a common alphabet $A = \{1, 2, \cdots, J\}$. We may then take Λ to be the $J-1$ dimensional set

$$\Lambda = \left\{ (\theta_1, \theta_2, \cdots, \theta_{J-1}) : \theta_j \geq 0, \sum_{j=1}^{J-1} \theta_j \leq 1 \right\}. \tag{41}$$

For each $\theta = (\theta_1, \theta_2, \cdots, \theta_{J-1}) \in \Lambda$, the probability $p_n(x|\theta)$ is given by

$$p_n(x|\theta) = \left\{ \prod_{j=1}^{J-1} \theta_j^{n_j(x)} \right\} \left(1 - \sum_{j=1}^{J-1} \theta_j \right)^{n_J(x)}, \tag{42}$$

where $n_j(x)$ is the number of occurrences of j in the vector x (i.e., $n_j(x)$ is the cardinality of $\{i: x_i = j\}$). This is of course just the familiar multinomial distribution. Let $\hat{\Theta} = (\hat{\Theta}_1, \hat{\Theta}_2, \cdots, \hat{\Theta}_{J-1})$ be the maximum likelihood estimate

$$\hat{\Theta}_j = g_j(X) \triangleq n^{-1} n_j(X), \tag{43}$$

so that

$$\operatorname{var}\{\hat{\Theta}_j | \Theta = \theta\} = n^{-1} \theta_j (1 - \theta_j). \tag{44}$$

Since $E\{\hat{\Theta}_j | \Theta = \theta\} = \theta_j$, then

$$\begin{aligned} D = E\{\|\hat{\Theta} - \Theta\|^2\} &= \int_\Lambda E\{\|\hat{\Theta} - \Theta\|^2 | \Theta = \theta\} w(\theta)\, d\theta \\ &= \int_\Lambda \sum_{j=1}^{J-1} E\{(\hat{\Theta}_j - \theta_j)^2 | \Theta = \theta\} w(\theta)\, d\theta \\ &= \frac{1}{n} \int_\Lambda \left\{ \sum_{j=1}^{J-1} \theta_j (1 - \theta_j) \right\} w(\theta)\, d\theta. \end{aligned} \tag{45}$$

If w is the uniform density on Λ (i.e., $w(\theta) = (J-1)!$ for all $\theta \in \Lambda$) then a simple calculation gives

$$D = \frac{(J-1)^2}{nJ(J+1)}. \tag{46}$$

Also, $h(w) = -\log[(J-1)!]$ for the uniform density and so the bound becomes

$$\mathcal{R}_n^* \geq \tfrac{1}{2}(J-1)\log n - \log[(J-1)!] - \tfrac{1}{2}(J-1)\log\left\{ \frac{2\pi e (J-1)^2}{J(J+1)} \right\}. \tag{47}$$

For example if $J = 2$

$$\mathcal{R}_n^* \geq \tfrac{1}{2} \log n - \tfrac{1}{2} \log(\pi e/3), \tag{48}$$

and if $J = 3$

$$\mathcal{R}_n^* \geq \log n - \log(4\pi e/3). \tag{49}$$

It is possible to improve the constant terms in (47)–(49) by using a more complicated density function w, but as we shall see the leading term $\frac{1}{2}(J-1)\log n$ is the best possible.

B. Upper Bounds

We now turn to upper bounds on $\mathcal{R}_n^*$. Suppose that q is a discrete density function on A^n such that for some $x_0 \in A^n$,

$$p_n(x|\theta) \leq q(x)/q(x_0), \qquad \text{for all } x \in A^n, \theta \in \Lambda. \tag{50}$$

If we let b_n be the Shannon code for q, then the codeword lengths for b_n are given by

$$l_n(x) = \lceil -\log q(x) \rceil.$$

The redundancy of this code is given by

$$\begin{aligned} r_n(l_n, \theta) &= n^{-1} \sum_x p_n(x|\theta) l_n(x) - H_n(\theta) \\ &\leq n^{-1} \left[1 - \sum_x p_n(x|\theta) \log q(x) \right] - H_n(\theta) \\ &= n^{-1} \left\{ 1 + \sum_x p_n(x|\theta) \log[p_n(x|\theta)/q(x)] \right\}. \end{aligned} \tag{51}$$

By assumption (50), (51) implies

$$r_n(l_n, \theta) \leq n^{-1}[1 - \log q(x_0)]. \tag{52}$$

We now apply this to the discrete memoryless source. Let q be the "mixture" distribution

$$q(x)=\int_{\Lambda} p_n(x|\theta)w(\theta)\,d\theta, \tag{53}$$

where $p_n(x|\theta)$ is the multinomial distribution of (42) and w is the Dirichlet density

$$w(\theta)=w(\theta_1,\theta_2,\cdots,\theta_{J-1})$$
$$=\Gamma(\tfrac{1}{2}J)\left[\Gamma(\tfrac{1}{2})\right]^{-J}\left\{\left(\prod_{j=1}^{J-1}\theta_j\right)\left(1-\sum_{j=1}^{J-1}\theta_j\right)\right\}^{-1/2}. \tag{54}$$

It follows that

$$q(x)=\Gamma(\tfrac{1}{2}J)\left\{\prod_{j=1}^{J}\Gamma\left(n_j+\tfrac{1}{2}\right)\right\}\Big/\left[\Gamma(\tfrac{1}{2})\right]^{J}\Gamma(n+\tfrac{1}{2}J), \tag{55}$$

where $n_j=n_j(x)$. Let x_0 be any constant vector (e.g., $x_0=(1,1,\cdots,1)$) so that

$$\frac{q(x)}{q(x_0)}=\frac{\prod_{j=1}^{J}\Gamma\left(n_j+\frac{1}{2}\right)}{\Gamma(n+\frac{1}{2})\left[\Gamma(\frac{1}{2})\right]^{J-1}}$$
$$=\frac{\prod_{j=1}^{J}(n_j+1)(n_j+2)\cdots(2n_j)}{(n+1)(n+2)\cdots(2n)}. \tag{56}$$

For a fixed x, the maximum value of $p(x|\theta)$ occurs for $\theta=(\theta_1,\theta_2,\cdots,\theta_J)$ such that $\theta_j=n^{-1}n_j(x)$; hence (50) is equivalent to the assertion

$$\prod_{j=1}^{J}(n_j/n)^{n_j}\leqslant\frac{\prod_{j=1}^{J}(n_j+1)(n_j+2)\cdots(2n_j)}{(n+1)(n+2)\cdots(2n)}. \tag{57}$$

Rearranged, this is

$$g_n(1)\leqslant\prod_{j=1}^{J}g_{n_j}(1), \tag{58}$$

where

$$g_k(u)=\prod_{i=1}^{k}(u+k^{-1}i). \tag{59}$$

The proof of the inequality in (58) is given in the Appendix.

It follows from (52) and

$$q(x_0)=\Gamma(\tfrac{1}{2}J)\Gamma(n+\tfrac{1}{2})/\Gamma(\tfrac{1}{2})\Gamma(n+\tfrac{1}{2}J),$$

that

$$\mathcal{R}_n^*\leqslant 1+\log\left\{\Gamma(\tfrac{1}{2})\Gamma(n+\tfrac{1}{2}J)/\Gamma(\tfrac{1}{2}J)\Gamma(n+\tfrac{1}{2})\right\}$$
$$=1+\log\left\{\prod_{i=0}^{n-1}(J+2i)(1+2i)^{-1}\right\}. \tag{60}$$

Using Stirling's asymptotic expansion of the gamma function, we get

$$\mathcal{R}_n^*\leqslant\tfrac{1}{2}(J-1)\log n-\log\left\{\Gamma(\tfrac{1}{2}J)/\Gamma(\tfrac{1}{2})\right\}+1+O(n^{-1}). \tag{61}$$

For example if $J=2$

$$n\mathcal{R}_n=\mathcal{R}_n^*\leqslant 1+\log\left\{\prod_{i=0}^{n-1}2(1+i)(1+2i)^{-1}\right\}$$
$$=\tfrac{1}{2}\log n+\tfrac{1}{2}\log\pi+1+O(n^{-1}), \tag{62}$$

and if $J=3$

$$n\mathcal{R}_n=\mathcal{R}_n^*\leqslant 1+\log\left\{\prod_{i=0}^{n-1}(3+2i)(1+2i)^{-1}\right\}$$
$$=\log n+2+O(n^{-1}). \tag{63}$$

In any case, (60) combined with (47) yields the asymptotic result

$$\mathcal{R}_n\sim\frac{(J-1)\log n}{2n}, \tag{64}$$

which is a generalization of a known asymptotic result [7] for binary sources ($J=2$).

IV. Construction of Universal Noiseless Source Codes

The proof of Lemma 1 and Theorem 1 provide a method for the construction of universal source codes. In this section, a code construction based on this method is given, and some preliminary results on its performance are presented. Modifications of the basic code construction are given in the next section along with more extensive numerical results on the performance of the various codes that are obtained from such modifications.

The basic technique for the code construction arises from a consideration of the class of binary memoryless sources. Furthermore, the general concepts are most easily explained in the context of memoryless sources. Consequently, we restrict attention in the present paper to the code construction and performance evaluation for binary memoryless sources. However it should be clear that our basic technique can be employed for M-ary memoryless sources, Markov sources, or more general sources with memory. Some preliminary results for Markov sources are given in [8].

Let the class of all binary memoryless sources be indexed by the parameter θ which takes values in the set Λ. Assume Λ is a subinterval of $[0,1]$. For most of our results $\Lambda=[0,1]$, but $\Lambda=[0.1,0.5]$ will also be considered. For each $\theta\in[0,1]$ let p_θ denote the discrete density function which is defined by $p_\theta(1)=\theta=1-p_\theta(0)$. Let ϵ be an arbitrary positive number and let $\delta=\epsilon/2$.

Since the sources in the class are memoryless, the discrete density functions $q^{(k)}$ of the proof of Lemma 1 can be taken to be "memoryless" as well; that is it suffices to

consider $q^{(k)}$ of the form

$$q^{(k)}(\boldsymbol{x}) = \prod_{i=1}^{n} p_{\varphi_k}(x_i)$$

for some choice of $\varphi_k \in \Lambda$. Consequently, the nth order relative entropy of (15) is equal to the first-order relative entropy. In all that follows we denote $H_1(p_\theta, p_\varphi)$ by $H(\theta, \varphi)$ for $\theta \in [0,1]$ and $\varphi \in (0,1)$. The goal is to select $\varphi_1, \varphi_2, \cdots, \varphi_k$ such that (c.f. (15))

$$H(\theta, \varphi_k) \leq \delta, \qquad \text{for all } \theta \in \Lambda^{(k)},$$

for each k where $\Lambda = \Lambda_1 \cup \Lambda_2 \cup \cdots \cup \Lambda_K$. Furthermore, the sets Λ_k can be taken to be intervals $[\theta_{k-1}, \theta_k]$ and it suffices to consider only $\varphi_k \in \Lambda_k$.

The following algorithm yields the set $\{\theta_k: 0 \leq k \leq K\}$ of endpoints of the intervals and the set $\{\varphi_k: 1 \leq k \leq K\}$ of code design probabilities. Let δ be an arbitrary positive number as in Lemma 1. The *first step* is to set $\theta_0 = 0$ and choose φ_1 such that

$$H(0, \varphi_1) = -\log(1 - \varphi_1) = \delta. \tag{65}$$

If $\varphi_1 < \frac{1}{2}$, then choose $\theta_1 > \varphi_1$ such that

$$H(\theta_1, \varphi_1) = \delta. \tag{66}$$

If $\theta_1 < \frac{1}{2}$ proceed to step two. At the *kth step* $\{\theta_i: 0 \leq i \leq k-1\}$ and $\{\varphi_i: 1 \leq i \leq k-1\}$ have already been selected, and $\varphi_k > \theta_{k-1}$ is selected to satisfy

$$H(\theta_{k-1}, \varphi_k) = \delta. \tag{67}$$

If $\varphi_k < \frac{1}{2}$, then choose $\theta_k > \varphi_k$ such that

$$H(\theta_k, \varphi_k) = \delta. \tag{68}$$

If $\theta_k < \frac{1}{2}$, proceed to step $k+1$. If at any time $\varphi_k \geq \frac{1}{2}$, then set $\varphi_k = \frac{1}{2}$, $K = 2k-1$, $\varphi_{K+1-i} = \varphi_i$ for $1 \leq i \leq k-1$, and $\theta_{K-i} = \theta_i$ for $0 \leq i \leq k-1$. If $\theta_k \geq \frac{1}{2}$, then set $\theta_k = \frac{1}{2}$, $k = 2k$, $\varphi_{K+1-i} = \varphi_i$ for $1 \leq i \leq k-1$, $\theta_{K-i} = \theta_i$ for $0 \leq i \leq k-1$.

The above procedure produces sets $\{\theta_k: 0 \leq k \leq K\}$ and $\{\varphi_k: 1 \leq k \leq K\}$ with the property that for each θ in the range $\theta_{k-1} \leq \theta \leq \theta_k$

$$H(\theta, \varphi_k) \leq \delta. \tag{69}$$

Consequently, for each $\theta \in [0,1]$ there exists at least one value of k for which (69) holds.

Let n be such that

$$n^{-1}(1 + \lceil \log K \rceil) \leq \delta, \tag{70}$$

and let $b_n^{(k)}$ be the nth order Shannon code for source φ_k so that

$$l_n^{(k)}(\boldsymbol{x}) = \lceil -\log p_n(\boldsymbol{x} | \varphi_k) \rceil. \tag{71}$$

Since $H_n(\theta) = H(\theta)$, then (17) implies $n^{-1}\bar{l}_n^{(k)}(\theta) = n^{-1}E_\theta\{l_n^{(k)}(\boldsymbol{X})\} \leq H(\theta) + \delta + n^{-1}$ for each $\theta \in \Lambda^{(k)}$. The codeword for a given source word $\boldsymbol{x}$ is defined as follows. Let k' be such that

$$l_n^{(k')}(\boldsymbol{x}) = \min\{l_n^{(k)}(\boldsymbol{x}): 1 \leq k \leq K\}, \tag{72}$$

and let β be the $\lceil \log K \rceil$ bit binary representation of the index k'. The codeword $b_n(\boldsymbol{x})$ for $\boldsymbol{x}$ is the prefix β followed by $b_n^{(k')}(\boldsymbol{x})$. Therefore for each $\boldsymbol{x}$ the length $l_n(\boldsymbol{x})$ of $b_n(\boldsymbol{x})$ is

$$l_n(\boldsymbol{x}) = \lceil \log K \rceil + \min\{l_n^{(k)}(\boldsymbol{x}): 1 \leq k \leq K\}. \tag{73}$$

Therefore the average codeword length for the code b_n satisfies

$$\begin{aligned} n^{-1}\bar{l}_n(\theta) &\leq n^{-1}\{\lceil \log K \rceil + \bar{l}_n^{(k)}(\theta)\} \\ &\leq n^{-1}\lceil \log K \rceil + H(\theta) + \delta + n^{-1} \\ &\leq H(\theta) + \epsilon \end{aligned}$$

for each $\theta \in \Lambda^{(k)}$. Since this holds for each k, the code b_n has average rate

$$n^{-1}\bar{l}_n(\theta) \leq H(\theta) + \epsilon, \qquad \text{for all } \theta \in \Lambda$$

which corresponds to average redundancy

$$r_n(l_n, \theta) \leq \epsilon, \qquad \text{for all } \theta \in \Lambda. \tag{74}$$

The above code was designed for $\Lambda = [0,1]$. Obvious modifications of the procedure gives a code for the case where Λ is any subinterval of $[0,1]$.

The procedure described above produces a set $\{\varphi_k: 1 \leq k \leq K\}$ of design probabilities for a given bound ϵ on the maximum redundancy. On the other hand, we can fix K and choose $\{\varphi_k: 1 \leq k \leq K\}$ to minimize the redundancy. For the latter approach we are interested in the quantities

$$\delta_K(\{\varphi_k\}, \Lambda) \triangleq \max\{\min\{H(\theta, \varphi_k): 1 \leq k \leq K\}: \theta \in \Lambda\} \tag{75}$$

and

$$\delta_K(\Lambda) \triangleq \min\{\delta(\{\varphi_k\}, \Lambda): \{\varphi_k: 1 \leq k \leq K\}\}. \tag{76}$$

$\delta_K(\Lambda)$ is the term in the upper bound on redundancy due to the relative entropy. The sets $\{\varphi_k: 1 \leq k \leq K\}$ which achieve $\delta_k(\Lambda)$ for $\Lambda = [0,1]$ have been determined numerically and are given in Table I. Because the effect of K on the redundancy bound is of the form $\lceil \log K \rceil$, values of $K \neq 2^i$ for some integer i are not included. To set the redundancy due to relative entropy below some δ, choose the minimum K such that $\delta_K(\Lambda) < \delta$ and use the corresponding $\{\varphi_k\}$.

Minimax redundancies achieved by this coding technique are given in Table II for various n and for the parameter sets $\Lambda = [0,1]$ and $[0.1, 0.5]$. Since most of the codewords are not used (only $1/K$ are used), the unused codewords were removed and the remaining ones shortened.

The lack of structure in these codes typically requires a table lookup scheme for decoding, so their complexity is a function of 2^n, the total number of codewords. This limits n, and thus the achievable redundancy is also limited. The problem of complexity may be alleviated by dividing the block of length n into n/n_s *subblocks* each of length n_s, and encoding these subblocks individually. Then the Shannon

TABLE I
OPTIMUM $\{\varphi_k: 1 \leq k \leq K\}$ FOR $\Lambda = [0,1]$

K	$\delta_K([0,1])$	$\{\varphi_k: 1 \leq k \leq K/2\}$ (others symmetric about $\theta = 0.5$)
2	0.3219	0.200
4	0.0935	0.063, 0.326
8	0.0254	0.018, 0.098, 0.234, 0.407
16	0.0067	0.005, 0.026, 0.065, 0.120, 0.189, 0.270, 0.358, 0.452

TABLE II
MAXIMUM REDUNDANCY FOR UNIVERSAL CODES OBTAINED FROM SHANNON CODES

n	$\Lambda = [0,1]$	$\Lambda = [0.1,0.5]$
5	0.400	0.195
8	0.375	0.166
10	0.300	0.154
15	0.200	0.122
20	0.150	0.099

TABLE III
MAXIMUM REDUNDANCY FOR UNIVERSAL CODES OBTAINED FROM SHANNON CODES

	$\Lambda = [0,1]$			$\Lambda = [0.1,0.5]$		
n_s	$n/n_s = 1$	3	10	$n/n_s = 1$	3	10
5	0.400	0.333	0.240	0.195	0.197	0.151
8	0.375	0.208	0.163	0.166	0.147	0.106
10	0.300	0.167	0.130	0.154	0.129	0.085
15	0.200	0.133	0.087	0.122	0.084	0.054
20	0.150	0.100	0.065	0.099	0.066	0.043

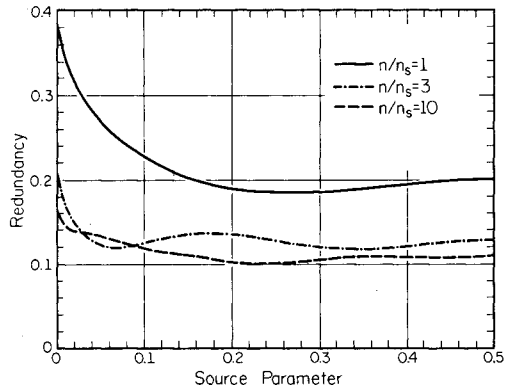

Fig. 1. Performance using Shannon codes for $n_s = 8$ and $n/n_s = 1$, 3, and 10.

codes for design probabilities remain as separate *subcodes*, and the system contains K parallel encoders. Each of these encoders generates a codeword for a block of length n by concatenating n/n_s codewords for subblocks of length n_s and a single prefix of length $\lceil \log K \rceil$. The shortest of the K codewords is then sent (cf. (21)). Whereas our previous bound on the redundancy was

$$r_n(l_n, \theta) \leq n^{-1}(1 + \lceil \log K \rceil) + \delta_K(\Lambda), \qquad (77)$$

for the subblock encoding scheme the bound becomes

$$r_n(l'_n, \theta) \leq n_s^{-1} + n^{-1}\lceil \log K \rceil + \delta_K(\Lambda). \qquad (78)$$

Although the redundancy bounds were derived for stationary sources, they remain valid for certain classes of nonstationary sources, including sources with slowly varying parameters. When subblocks are encoded, n is no longer limited by complexity but only by possible variations in source parameters as a function of time. If n becomes so large that the parameter is no longer approximately constant over one block length, then the bounds on redundancy will not apply. In general this constraint is much weaker than that imposed by complexity.

Table III compares performance for several values of n_s, n/n_s, and $\Lambda = [0,1]$ and $[0.1,0.5]$. Results for the original scheme ($n = n_s$) are included for comparison. Fig. 1 gives redundancy versus θ for the cases with $n_s = 8$ and $\Lambda = [0,1]$. Table IV compares the performance of different n_s for the same n. If n_s does not divide n, the actual block length of the code is given in parentheses (e.g., for $n_s = 8$ and $n \approx 100$, the actual block length n is 96). Fig. 2 and 3 give actual performance curves for $n \approx 100$ and $\Lambda = [0,1]$ and $[0.1,0.5]$. (Note that for $\Lambda = [0,1]$, code performance is symmetric about $\theta = 0.5$.)

TABLE IV
MAXIMUM REDUNDANCY FOR UNIVERSAL CODES OBTAINED FROM SHANNON CODES

	$\Lambda = [0,1]$			$\Lambda = [0.1,0.5]$		
n_s	$n = 10$	30	100	$n = 10$	30	100
5	0.400	0.267	0.230	0.229	0.164	0.150
8	0.375 (8)*	0.186 (32)	0.157 (96)	0.166 (8)	0.137 (32)	0.110 (96)
10	0.300	0.167	0.130	0.154	0.129	0.085
15		0.167	0.095 (105)		0.101	0.060 (105)
20			0.080			0.053

*The exact value of n is given in parentheses for those cases in which it differs from the column heading.

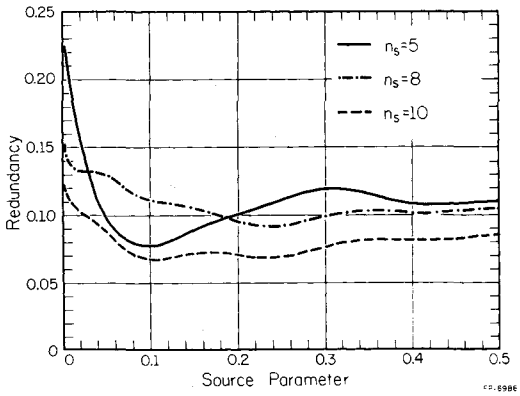

Fig. 2. Performance using Shannon codes for $\Lambda = [0,1]$, $n \approx 100$, and various n_s.

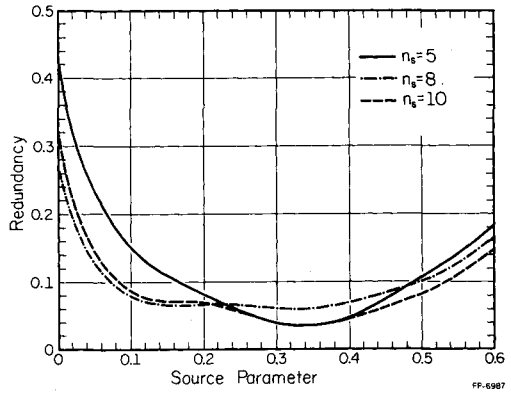

Fig. 3. Performance using Shannon codes for $\Lambda = [0.1,0.5]$, $n \approx 100$, and various n_s.

V. MODIFIED CONSTRUCTIONS AND ALTERNATIVE CODES

The maximum redundancy of our basic coding technique may be decreased by several simple modifications. First Huffman coding may be used instead of Shannon coding. Although the bound on average codeword length given by (17) was derived for Shannon codes and does not apply to Huffman codes (nor is it easily extendable), actual performance was found to be better for Huffman codes than for Shannon codes. However there is generally no improvement for $\theta = 0$ or 1 (the redundancy here being determined by the requirement of at least one bit for any codeword), and so in certain cases the maximum redundancy may be unaffected. In these cases the further modification of using a subblock length $n_s' > n_s$ for the subcodes designed for the ends of the interval $[0,1]$ can reduce the maximum redundancy considerably while reasonable complexity is maintained by providing codewords for only a limited number of the possible output blocks of length n_s'. (Note n_s' must divide n.) In many cases these special "long codes" consisted of a single codeword for all zeros (or all ones) block, in which case only a subcode prefix would be sent when n/n_s' of these blocks occurred. In other cases the "long code" had a codeword for all zeros (or all ones) block and codewords for blocks with a single one (or zero). So the number of codewords in the "long code" was at most $n_s' + 1$. Fig. 4 compares the performance of universal codes derived from Shannon codes, Huffman codes, and Huffman codes with a "long code" for each end of the interval $[0,1]$.

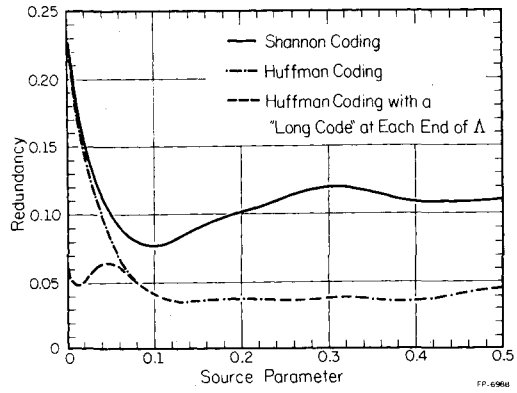

Fig. 4. Effect of modifications for $\Lambda = [0,1]$, $n = 100$, and $n_s = 5$.

In coding for $\Lambda = [0.1,0.5]$, the redundancy at $\theta = 0.1$ and 0.5 is higher than that over the interior of the set; therefore the maximum redundancy is decreased by moving the design probabilities closer to these values of θ (see Fig. 5). This repositioning of design probabilities is also useful in a number of other cases where one or two values of θ dominate the performance. Finally the performance is improved in some cases by use of variable-length prefixes for $K \neq 2^i$ (with shorter prefixes on the end subcodes; i.e., the "long codes"). Tables V and VI give the best results obtained using these modifications for the same n, n_s, and Λ as Tables III and IV. Fig. 6 gives best results for the same case as Fig. 2.

The results obtained with Shannon coding and the best results using the proposed modifications are compared in

TABLE V
BEST RESULTS FOR VARIOUS n_s, n/n_s, AND Λ

	Maximum Redundancy					
	$\Lambda=[0,1]$			$\Lambda=[0.1,0.5]$		
n_s	$n/n_s=1$	3	10	$n/n_s=1$	3	10
5	0.400	0.179	0.080	0.139	0.075	0.050
8	0.250	0.133	0.050	0.105	0.057	0.030
10	0.200	0.100	0.050	0.081	0.051	0.026
15	0.137	0.074	0.040	0.071	0.041	0.021
20	0.120	0.067	0.025	0.058	0.032	0.018

TABLE VI
BEST RESULTS FOR VARIOUS n_s, n, AND Λ

	Maximum Redundancy					
	$\Lambda=[0,1]$			$\Lambda=[0.1,0.5]$		
n_s	$n=10$	30	100	$n=10$	30	100
5	0.200	0.106	0.70	0.103	0.051	0.033
8	0.250 (8)*	0.094 (32)	0.052 (96)	0.105 (8)	0.051 (32)	0.028 (96)
10	0.200	0.100	0.050	0.081	0.051	0.026
15		0.100	0.043 (105)		0.047	0.024 (105)
20			0.040			0.024

*The exact value of n is given in parentheses for those cases in which it differs from the column heading.

TABLE VII
COMPARISON OF CODE PERFORMANCE WITH BOUNDS ON nTH ORDER MINIMAX REDUNDANCY

	Maximum Redundancy			
n	Lower Bound	Upper Bound	Shannon Coding	Best Codes
5	0.081	0.605	0.400	0.400
8	0.093	0.419	0.375	0.250
10	0.091	0.351	0.300	0.200
15	0.080	0.253	0.200	0.137
20	0.070	0.200	0.150	0.120
30	0.057	0.143	0.167 (10)*	0.100 (10)
100	0.026	0.031	0.080 (20)	0.040 (20)
200	0.015	0.016	0.065 (20)	0.025 (20)

*If subblock encoding was employed, the subblock length n_s is shown in parentheses.

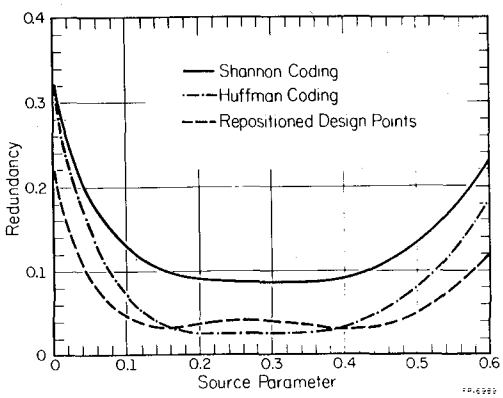

Fig. 5. Effect of modifications for $\Lambda=[0.1,0.5]$, $n=100$, and $n_s=5$.

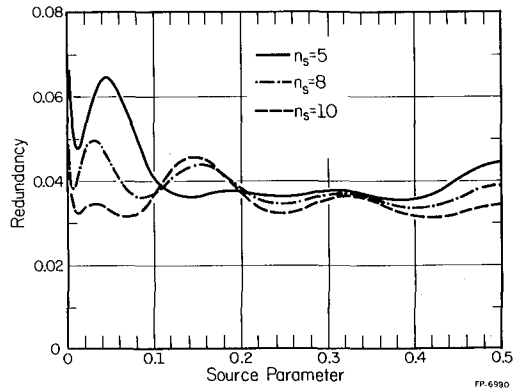

Fig. 6. Best results for $\Lambda=[0.1]$, $n\approx 100$, and various n_s.

Table VII with the bounds derived in Section III. The code performance for $n \geq 30$ applies to the encoding of subblocks of length n_s. The value of n_s is shown in parentheses next to the value of the redundancy.

Another code that is strongly universal for the class of binary memoryless sources is the Shannon code for the mixture distribution given by

$$g(x)=\left\{(n+1)\binom{n}{k}\right\}^{-1} \tag{79}$$

for all x such that $\mathcal{U}(x)=k$, where $\mathcal{U}(x)=\sum_{i=1}^{n} x_i$ is the number of ones in the vector x. This mixture distribution,

which was previously employed in [1], is obtained from (53) by letting w be the uniform density on $\Lambda=[0,1]$. Notice that if $\boldsymbol{x}_0=(0,0,\cdots,0)$ then $q(\boldsymbol{x}_0)=(n+1)^{-1}$ and for $\boldsymbol{x}$ such that $\mathcal{N}(\boldsymbol{x})=k$

$$p_n(\boldsymbol{x}|\theta)=\theta^k(1-\theta)^{n-k}\leqslant\binom{n}{k}^{-1}=\frac{q(\boldsymbol{x})}{q(\boldsymbol{x}_0)}. \quad (80)$$

The inequality in (80) follows from the fact that $\binom{n}{k}\theta^k(1-\theta)^{n-k}$, $0\leqslant k\leqslant n$, is a probability distribution. Since (50) is satisfied, we can conclude from (52) that the Shannon code for q has redundancy

$$r_n(l_n,\theta)\leqslant n^{-1}[1+\log(n+1)]. \quad (81)$$

It follows that

$$\mathcal{R}_n\leqslant n^{-1}[1+\log(n+1)] \quad (82)$$

which is not as tight as (62).

Suppose that instead of coding with respect to the mixture probability distribution (79) we use a two part codeword which consists of a fixed-length part that specifies $\mathcal{N}(\boldsymbol{x})$ the number of ones and a variable-length part that specifies the location of the ones. This encoding method is considered in several papers such as [1, pp. 786–787] and [2]. For this code the redundancy is given by

$$r_n(l_n,\theta)=n^{-1}\sum_{k=0}^{n}\left\{\lceil\log(n+1)\rceil+\left\lceil\log\binom{n}{k}\right\rceil\right\}\cdot\binom{n}{k}\theta^k(1-\theta)^{n-k}-\mathcal{H}_b(\theta) \quad (83)$$

$$\leqslant n^{-1}\{1+\lceil\log(n+1)\rceil-H(S_n)\} \quad (84)$$

where S_n is a random variable which has the binomial distribution with parameters n and θ. Since $H(S_n)\geqslant 0$ we have that

$$r_n(l_n,\theta)\leqslant n^{-1}\{1+\lceil\log(n+1)\rceil\}. \quad (85)$$

It follows from (85) that

$$\mathcal{R}_n\leqslant n^{-1}\{1+\lceil\log(n+1)\rceil\}, \quad (86)$$

which is not quite as tight as (82). The primary interest in the coding method described above is due to the fact that the encoding and decoding can be implemented without use of a table lookup procedure. However, notice that (83) implies

$$\hat{r}_n(l_n)\geqslant n^{-1}\lceil\log(n+1)\rceil. \quad (87)$$

A comparison of (87) with the data in Table VII shows that the code constructions that we have presented in this paper yield considerably smaller redundancies for small to intermediate block lengths.

APPENDIX

Lemma: For $n=0,1,2,\cdots$, define

$$g_n(x)=\prod_{k=1}^{n}\left(x+\frac{k}{n}\right).$$

Then if $n=n_1+n_2+\cdots+n_r$, and if $x\geqslant 0$,

$$g_n(x)\leqslant\prod_{i=1}^{r}g_{n_i}(x).$$

Proof: We prove the case $r=2$; the general result follows immediately by induction. It is required to show that for nonnegative integers n, m, and $x\geqslant 0$,

$$\prod_{i=1}^{n+m}\left(x+\frac{i}{n+m}\right)\leqslant\prod_{j=1}^{n}\left(x+\frac{j}{n}\right)\cdot\prod_{k=1}^{m}\left(x+\frac{k}{m}\right). \quad (A.1)$$

There are $n+m$ terms on both sides of (A.1). We will show that it is possible to put them in one-to-one correspondence in such a way that each term on the left will be less than or equal to its corresponding term on the right.

Thus denote by S_i the set of terms on the right side of (A.1) which are greater than or equal to the term $x+i/(n+m)$ which appears on the left side of (A.1). Our proof will be complete if we can show that there exists an ordering $(t_1,t_2,\cdots,t_{n+m})$ of the $n+m$ terms on the right side of (A.1) such that $t_i\in S_i$, $i=1,2,\cdots,m+n$.

Now the term $x+i/(n+m)$ is less than or equal to the term $x+j/n$ if and only if $j\geqslant ni/(n+m)$; similarly it is less than or equal to $x+k/m$ if and only if $k\geqslant mi/(n+m)$. Hence the number of terms in the set S_i is $(n-\lceil ni/(n+m)\rceil+1)+(m-\lceil mi/(n+m)\rceil+1)$. Using the simple inequality $\lceil x\rceil+\lceil y\rceil\leqslant\lceil x+y\rceil+1$, it follows that

$$|S_i|\geqslant n+m+1-i. \quad (A.2)$$

This inequality allows us to exhibit the desired ordering of the terms on the right side of (A.1). Simply choose $t_{n+m}\in S_{n+m}$ (by (A.2) S_{n+m} is not empty), and having chosen $t_{n+m},t_{n+m-1},\cdots,t_{i+1}$, choose $t_i\in S_i$. This is always possible since at most $n+m-i$ terms from S_i have already been chosen.

REFERENCES

[1] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[2] V. F. Babkin, "A universal encoding method with a nonexponential work expenditure for a source of independent messages," *Probl. Peredach. Inform.* vol. 7, no. 4, pp. 13–21, 1971 (English translation in *Prob. Inform. Transmiss.*, vol. 7, no. 4, pp. 288–294, Jan. 1974).

[3] B. M. Fitingof, "Optimal coding in the case of unknown and changing message statistics," *Prob. Peredach. Inform.* vol. 2, no. 2, pp. 3–11, 1966 (English translation in *Prob. Inform. Transmiss.*, vol. 2, no. 2, pp. 1–7, 1968).

[4] ——, "The compression of discrete information," *Probl. Peredach. Inform.* vol. 3, no. 3, pp. 28–36, 1967 (English translation in *Prob. Inform. Transmiss.*, vol. 3, no. 3, pp. 22–29, 1969).

[5] E. N. Gilbert, "Codes based on inaccurate source probabilities," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 304–314, May 1971.

[6] K. M. Mackenthun and M. B. Pursley, "Variable-rate universal block source coding subject to a fidelity constraint," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 349–360, May 1978.

[7] R. E. Krichevskii, "The relation between redundancy coding and the reliability of information from a source," *Prob. Peredach. Inform.*, vol. 4, no. 3, pp. 48–57, 1968 (English translation in *Prob. Inform. Transmiss.*, vol. 4, no. 3, pp. 37–45, 1971).

[8] R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Universal noiseless data compression for Markov sources," in *Proc. 1980 Conf. Inform. Sci. and Syst.*, Princeton Univ., pp. 164–167, Mar. 1980.

[9] M. B. Pursley and L. D. Davisson, "Variable rate coding for nonergodic sources and classes of ergodic sources subject to a fidelity constraint," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 324–337, May 1976.

[10] ——, "Mismatch bounds for variable-rate source codes with applications to universal data compression," in *Proc. AFOSR Workshop in Commun. Theory and Appl.*, Provincetown, MA, pp. 33–37, Sept. 1978.